**The Association of Dengue Disease with Temperature, Precipitation and Vegetation index in Tropical and Sub-tropical Parts of the World**

**Introduction to the Research Question**

The purpose of this study was to predict the number of dengue cases each week for each location (city, year and week of year in San Juan and Iquitos) based on environmental variables and find best predictor factors describing how changes in maximum, minimum and average air temperatures, total precipitation, relative and specific humidity and satellite measured vegetation index affect number of disease cases.

As data analyst, my current task is to predict the number of dengue cases each week (in each location) based on environmental variables describing changes in temperature, precipitation, vegetation and other factors.

The research to identify factors associated to epidemic diseases is very important to the public health and may help to better understand if in these case factors are related to climate change. These days many of the nearly half billion dengue cases per year occurring in Latin America. In many cases dengue disease causes severe health problems and even death. Accurate dengue predictions would help public health workers and people around the world take steps to reduce the impact of these epidemics.  Although the relationship with climate is complex, a growing number of scientists argue that climate change is likely to produce distributional shifts that will have significant public health implications worldwide.

**Methods**

**Sample**

The sample included N=1111 weekly environmental measurements for San Juan (Puerto Rico) tropical city and Iquitos (Peru) sub-tropical cities. Data, indicators and measurement provided by NOAA's GHCN ([daily climate data](https://www.ncdc.noaa.gov/oa/climate/ghcn-daily/) weather station measurements), NOAA's NCEP ([Climate Forecast System Reanalysis](http://rda.ucar.edu/datasets/ds093.0/#metadata/detailed.html?_do=y) measurements), NOAA's [CDR (Normalized Difference Vegetation Index](https://www.ncdc.noaa.gov/cdr)), PERSIANN [satellite precipitation measurements](http://www.ncdc.noaa.gov/cdr/operationalcdrs.html).

**Measures**

The response variable TOTAL\_CASES represents weekly counts of dengue cases for up to 52 weeks per year for each location, San Juan (SJ) and Iquitos (IQ) from 1990 to 2010.

Predictors for San Juan (SJ) and Iquitos (IQ) included:

1. Week of the year (weekofyear), quantitative variable, week id’s ranging from 1 to 52
2. Mean of specific humidity (specific\_humidity\_g\_per\_kg), quantitative variable
3. Total millimeters precipitation amount (station\_precip\_mm)
4. Satellite average vegetation index (Satellite average vegetation index (vegitation\_index\_avg)

A**nalysis**.

The distribution of dengue cases and all predictors were evaluated by examining mean, standard deviation and minimum and maximum values for all quantitative variables, including univariate analysis for outliners. Based on analysis results, outliners were kept in dataset.

The Pearson correlation was used to test correlation coefficient between variables. The General linear model (GLM) was used to test basic linear regression model for the association between explanatory variables and response variable to test strength of relationship between variables. The multiple regression including STEPWISE variable selection was also used to test possible relationship between primary variable and additional confounders (variables) in the model, various graphs and plots were also used for analysis of data distribution.

To predict total dengue cases Penalized regression method (Lasso - Least Absolute Selection and Shrinkage Operator) regression was used to test model and to provide greater prediction accuracy for both locations (N=1111) and for each separately, San Juan N=724 and Iquitos N=387. Prior to conducting LASSO regression all predictor variables were standardized with mean=0 and standard deviation=1. The estimation of LASSO regression model was performed with 70% of training set and 30% of test set for both and each location separately.

In addition, K-MEANS cluster analysis were applied on training set to create K=1-10 clusters using Euclidean distance to partition observations into smaller set of clusters based on similarity of responses on multiple variables. All clustering variables were standardized using STANDARD procedure to have also a mean of 0 and standard deviation of 1. The training and test sets created with 70% in training and 30% in test. Observations with missing values removed prior creation of both sets. Iquitos (IQ) training set N=271), test N=116, San Juan (SJ) training set N=507, test N=217.